CELPIP Test Review Report II : Psychometric Properties

Test Research and Development

Paragon Testing Enterprises

April 2021

CELPIP Test Review Report II :

Psychometric Properties

Test Research and Development

Paragon Testing Enterprises

April 2021

Citation:
Paragon Testing Enterprises. (2021, April). *CELPIP Test Review Report II: Psychometric Properties* (No. 2021–2). Vancouver, BC: Paragon Testing Enterprises.

Table of Contents

List of Tables

List of Figures

1. Introduction

The purpose of this report is to present the psychometric properties of the Canadian English Language Proficiency Index Program (CELPIP) Test to demonstrate that the test produces consistently similar scores among candidates with similar language proficiency, and that this consistency applies across different versions and multiple administrations of the test over time, as well as across gender and language groups. In addition, the psychometric properties of the test suggest that it distinguishes consistently and accurately between levels of language proficiency relevant to the Canadian Language Benchmarks as evidenced by established methods of analysis of scored responses. The analyses presented in this report provide evidence based on the *internal structure* (*Test Standards*, AERA, APA, & NCME, 2014) of the CELPIP Test to support its proposed score use and interpretation.

The CELPIP Test is a complete English language testing program designed to measure the functional language proficiency required for successful communication in general Canadian social, educational, and workplace contexts. The test is completely computer-delivered and can be taken in one of Paragon's designated test centres at computer stations. The CELPIP-General Test consists of four components: listening, reading, writing, speaking. The CELPIP-General LS Test consists of two of the four components: listening and speaking, with the same structure and specifications of the listening and speaking components of the CELPIP-General Test. The listening and reading components of the test are computer-scored, while the writing and speaking components are recorded and rated via an online rating system by trained raters located across Canada.

Paragon Testing Enterprises has a Service Agreement with the Government of Canada to deliver language testing services for economic immigration. CELPIP has been designated by the Immigration, Refugees and Citizenship Canada (IRCC) as evidence of English language proficiency for applications for permanent residence immigration in Canada and Canadian citizenship.

The report is organized as follows. Section 2 describes the test-taker data used for the analyses in this report. Section 3 provides a summary of the psychometric properties of test forms for the listening (3.1), reading (3.2), writing (3.3), and speaking (3.4) components of CELPIP. Section 4 reports the scoring accuracy of the test, including scoring accuracy and consistency across test-centre locations (Section 4.1), by scoring band (4.2), and the conditional standard error of measurement by test components (4.3). Section 5 describes the correlations among the four abilities, listening, reading, writing and speaking. Section 6 presents the results of differential item functioning analysis, including gender DIF (6.1) and language group DIF (6.2). Finally, Section 6.3 reports on test drift analysis, including item parameter drift over time (6.3.1) and rating drift (6.3.2).

Besides the analyses presented in this report, Paragon also implements a number of operational quality control procedures to ensure that the CELPIP Test produces accurate and consistent test scores across the multiple parameters discussed above, including rigorous item review and tryout procedures and ongoing rater training and feedback. For more information on CELPIP test structure, content, quality control and item development procedures, please see CELPIP Test Review Report I.

2. Summary of Data

The CELPIP Test is officially accepted by several governments, professional organizations, colleges, universities, and employers. The test-taker population reported in this report includes three years of CELPIP administrations between March 1, 2017 and March 1, 2020. This includes 211,661 listening and speaking test takers, and 165,045 reading and writing test takers. One date range is chosen for all of the information presented in this report, which allows Paragon to use one date throughout the Report with the aim of simplifying the report for readers. The reported time range for the data has been selected to reflect typical test-taker performance in the state prior to the Covid-19 pandemic. The same data set is used for all analyses in this report, which includes all test takers who have completed all components in a test session (CELPIP-General: 4 components; CELPIP-General LS: 2 components).

This sample consists of 44.6% female and 55.4% male test takers. Table 1 displays the age-group distributions. The top 10 self-declared first language groups represented in this sample are: English, Spanish, Chinese, Tagalog, Arabic, Portuguese, Panjabi, Hindi, Korean, and Farsi. Together, these language groups represent 67.7% of the test takers.

Table 1. Distribution of the Age Groups

AGE GROUP (YEARS)	PERCENTAGE OF SAMPLE
<= 20	0.0
21-25	5.6
26-30	24.2
31-35	26.7
36-40	20.1
41-45	11.8
46-50	6.7
> 50	4.9
TOTAL	100.0

3. Summary of the Psychometric Properties of the Test Forms

In the following, the psychometric properties of test forms for listening, reading, writing, and speaking components of CELPIP are summarized. For listening and reading test forms, classical test theory (CTT) and item response theory (IRT) statistics at the form level are presented (Table 2 and Table 3). For writing and speaking forms, classical test theory statistics are reported (Table 4 and Table 5).

To ensure test security, Paragon maintains a large item pool which generates thousands of test forms. Due to the large number of test forms, many forms have only been administered to a small number of test takers, making their statistics hard to interpret. For reporting purpose, the CTT statistics were summarized for forms with a minimum of 200 test takers, and the IRT statistics for listening and reading components were presented for the most common 20 forms (the minimum sample size for listening forms is 763 and 619 for reading forms).

3.1. Listening Test Forms

Table 2 summarizes the CTT-based form properties. A total of 166 listening forms meet the minimum sample size of 200 for reporting the CTT-based psychometric properties. The columns list a range of psychometric properties, including reliability (measured by internal consistency, i.e., Cronbach's alpha) and descriptive statistics of raw score distributions (mean, standard deviation, median, interquartile range, skewness, kurtosis, minimum, and maximum scores). Taken together, these statistics depict the internal consistency of test items and raw score distributions of test takers. The analysis was run for each of the 166 listening forms. To summarize the psychometric properties of all these forms, the range (i.e., minimum and maximum), mean, and variation (i.e., standard deviation) of these form-level statistics are described in Table 2.

Table 2. Summary Statistics of Listening Forms* (N=166)

		RAW SCORE DISTRIBUTIONS ACROSS TEST FORMS**							
SUM. STATS.	**INTER. CONS.**	**Mean**	**SD**	**Median**	**IQR**	**Skew**	**Kurtosis**	**Min**	**Max**
MIN	0.84	24.43	6.01	25.00	7.00	-1.43	-0.89	0.00	37.00
MAX	0.92	30.00	8.25	32.00	13.00	-0.30	1.99	11.00	38.00
MEAN	0.89	27.49	7.13	29.04	9.96	-0.80	-0.01	5.11	37.98
SD	0.01	1.18	0.46	1.46	1.31	0.20	0.51	1.88	0.15

Note: Sum. Stat. = summary statistics, Inter. Cons. = internal consistency, Min = minimum, Max = maximum, SD = standard deviation, IQR = interquartile range. * This table reported on the test forms that have been administered to more than 200 test takers during the reporting time period. ** For the CELPIP listening component, the final scores are reported on the 11-point reporting scale (M, 3-12) which are converted from equated true scores rather than raw scores.

As shown in Table 2 above, for the listening component, the reliability for each form is at least 0.84 (average = 0.89), and the mean scores on each test form (i.e., average number-correct scores) are between 24.43 and 30. The interquartile range (IQR) illustrates the difference between the first and third quartile of an ordered range of data, which measures the spread of data. Raw scores for all listening forms are negatively skewed, meaning that more test takers tend to score higher on the listening forms in the raw score metric.

The CTT-based psychometric properties were calculated based on raw responses. By applying the IRT model to calibrate items and equate test scores, some of these form-to-form differences are accounted for and do not significantly affect test-takers' CELPIP scores.

A 2PL IRT model was fitted to the CELPIP Test listening component and the item parameters, *a* and *b*, are estimated by the IRT software, flexMIRT (Cai, 2017). Figure 1 below provides a graphical presentation of the standard error of measurement for the 20 listening test forms with the largest sample sizes.

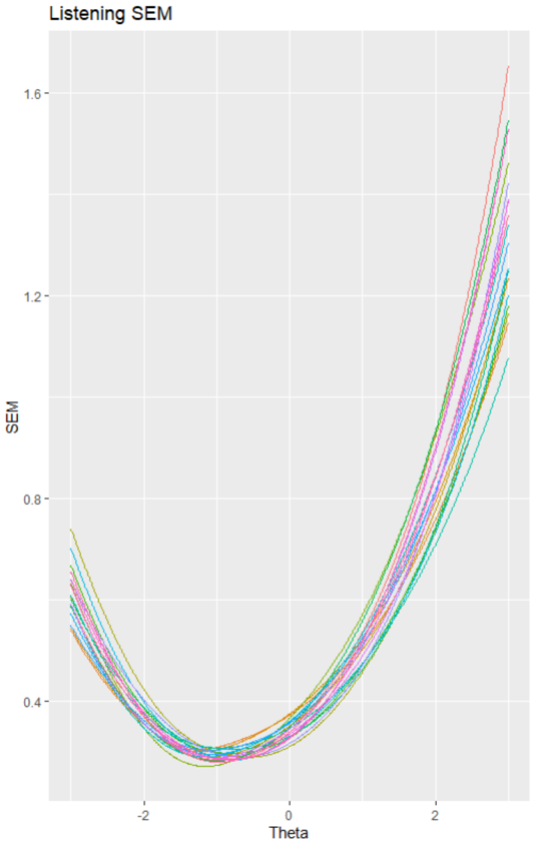

Figure 1. Standard Error of Measurement for 20 CELPIP Listening Test Forms

3.2. Reading Test Forms

Similar to listening forms, CTT statistics for reading forms with at least 200 test takers are reported. As shown in Table 3 below, for the reading component, reliability for each form is at least 0.84 (average = 0.89), mean scores on each test form (i.e., average number-correct scores) are between 22.61 and 29.45. The interquartile range (IQR) shows the spread of data. Raw scores for most reading forms are negatively skewed, indicating that more test takers tend to score higher on the reading forms in the raw score metric.

Table 3. Summary Statistics of Reading Forms* (N=165)

		RAW SCORE DISTRIBUTIONS ACROSS TEST FORMS**							
SUM. STAT.	**INTER. CONS.**	Mean	SD	Median	IQR	Skew	Kurtosis	Min	Max
MIN	0.84	22.61	6.16	23.00	8.00	-1.21	-1.18	0.00	37.00
MAX	0.93	29.45	9.13	31.00	17.00	0.04	0.98	11.00	38.00
MEAN	0.89	26.24	7.20	27.22	10.45	-0.55	-0.34	5.58	37.96
SD	0.02	1.20	0.51	1.55	1.28	0.20	0.36	1.94	0.19

Note: Sum. Stat. = summary statistics, Inter. Cons. = internal consistency, Min = minimum, Max = maximum, SD = standard deviation, IQR = interquartile range. * This table reported on the test forms that have been administered to more than 200 test takers during the reporting time period. ** For the CELPIP Test reading component, the final scores are reported on the 11-point reporting scale (M, 3-12) which are converted from equated true scores rather than raw scores.

A 2PL IRT model was fitted to the CELPIP Test reading component and the item parameters, *a* and *b*, are estimated by the IRT software, flexMIRT (Cai, 2017). Figure 2 below provides a graphical presentation of the standard error of measurement for the 20 reading test forms with the largest sample sizes.

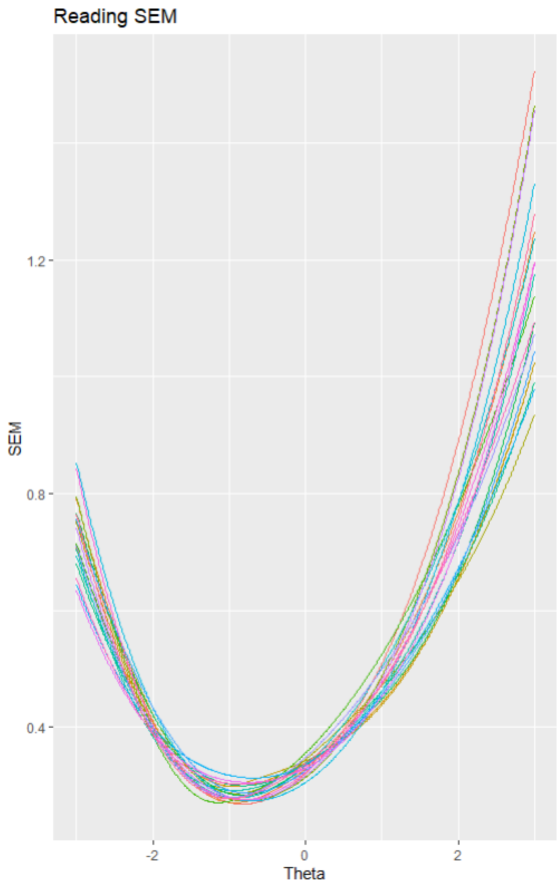

Figure 2. Standard Error of Measurement for 20 CELPIP Reading Test Forms

3.3. Writing Test Forms

CTT statistics are reported for 215 writing forms, all of which had a sample size of 200 or larger. As shown in Table 4 below, for the writing component, reliability for each form is at least 0.88 (average = 0.92), mean scores on each test form are between 6.78 and 8.18. The interquartile range (IQR) shows the spread of data. Skewness and Kurtosis illustrate the overall score distributions.

Table 4. Summary Statistics of Writing Forms (N=215)*

		RATING DISTRIBUTIONS ACROSS TEST FORMS							
SUM. STAT.	**INTER. CONS.**	Mean	SD	Median	IQR	Skew	Kurtosis	Min	Max
MIN	0.88	6.78	0.99	6.86	1.09	-1.32	-0.33	0.00	9.35
MAX	0.94	8.18	2.08	7.90	2.89	0.53	5.42	4.38	12.06
MEAN	0.92	7.12	1.29	7.29	1.45	-0.26	1.20	2.36	11.22
SD	0.01	0.20	0.15	0.16	0.20	0.28	0.67	0.99	0.58

Note: Sum. Stat. = summary statistics, Inter. Cons. = internal consistency, Min=minimum, Max=maximum, SD=standard deviation, IQR= interquartile range. * This table reported on the test forms that have been administered to more than 200 test takers during the reporting time period.

3.4. Speaking Test Forms

For speaking component, CTT statistics based on 162 forms show that, at the form level, the reliability is at least 0.95 (average = 0.97). Mean scores on each test form are between 6.39 and 7.74. The Skewness and Kurtosis illustrate the overall score distributions, showing the speaking ratings are slightly positivity skewed.

Table 5. Summary Statistics of Speaking Forms (N=162)*

		RATING DISTRIBUTIONS ACROSS TEST FORMS							
SUM. STAT.	**INTER. CONS.**	Mean	SD	Median	IQR	Skew	Kurtosis	Min	Max
MIN	0.95	6.39	1.66	6.20	1.74	0.19	-0.79	0.00	11.77
MAX	0.98	7.74	2.41	7.31	3.57	1.02	1.32	3.38	11.86
MEAN	0.97	7.06	2.03	6.68	2.36	0.59	0.14	1.31	11.86
SD	0.00	0.21	0.13	0.20	0.27	0.13	0.38	1.02	0.01

Note: Sum. Stat. = summary statistics, Inter. Cons. = internal consistency, Min = minimum, Max = maximum, SD = standard deviation, IQR = interquartile range. * This table reported on the test forms that have been administered to more than 200 test takers during the reporting time period.

4. Summary of Scoring Accuracy

All the responses from test takers at different geographic locations are collected and stored in Paragon’s database. Rating assignments are then created and randomly distributed to active raters without disclosing test-takers' personal information (including where they took the test). Thus, raters are unaware of test-takers' personal details and, due to random assignment, rating assignments are not dependent on the location of test centres or raters. Neither the test centre nor the rater location is a systematic source of variance in test-takers’ scores.

The quality of the writing and speaking scores depends on the ability of raters to agree on their ratings. To ensure scoring agreement, between and within raters, different analytic techniques were employed, including exact/adjacent agreement analysis, rater calibration exercises, and multi-faceted Rasch modelling (MFRM). Paragon’s Psychometrics team generates weekly reports based on these analyses for the CELPIP rating team so that they could provide raters with feedback on their performance.

4.1. Scoring Accuracy and Consistency across Test Location

For demonstrative purposes only, Table 6 presents the standard error of test-takers’ component scores for tests taken in Canadian (domestic) and international test centres. These statistics show the average variability of an individual test-taker’s item/task scores for a given component. Overall, the standard errors of the scores are similar between domestic and international test centres. Note that the numbers of test takers who took the test in different locations differ greatly. To ensure the interpretability and accuracy of the statistics, only the locations with more than 100 test takers over the 3-year reporting period are reported.

Table 6. Average Standard Error of Scores: Domestic vs. International

REGION	LISTENING	READING	WRITING	SPEAKING
DOMESTIC (CANADA)	0.29	0.30	0.27	0.27
INTERNATIONAL	0.26	0.29	0.26	0.27

When test takers are classified into different proficiency levels based on their test performance, it is important to evaluate the degree to which the classifications are accurate and consistent. In the following sections, results on scoring accuracy and consistency are presented for CELPIP levels 4 through 10, which correspond to CLB 4 through 10 (these are the levels that are perceived as the most relevant by many score users such as the IRCC). Three measures of scoring consistency and accuracy are provided for each band level, including decision accuracy (Section 4.2), decision consistency (Section 4.2), and standard error of measurement (Section 4.3). These measures provide supporting evidence for the accuracy and consistency of CELPIP test scores.

4.2. Decision Consistency and Decision Accuracy by Scoring Band

Livingston and Lewis (1995) defined decision accuracy (DA) or classification accuracy as the “extent to which the actual classifications of the test takers agree with those that would be made on the basis of their true score, if their true scores could somehow be known” (p. 180).

To be consistent with the scoring method employed for the four components of the CELPIP Test, Paragon employs Rudner’s method (Rudner, 2001, 2005) to calculate the decision consistency and accuracy for the listening and reading components using the R package cacIRT (Lathrop, 2014). The Livingston and Lewis Method (1995) is adopted to estimate the decision consistency and accuracy of the speaking and writing components, using the computer program BB-Class (Brennan, 2004).

For each band level, four statistics are reported in Table 7– Table 10. Decision consistency measures the proportion of agreement in classification of scores between two test forms of the same difficulty. Decision accuracy estimates the probability of consistent classifications across true scores and observed scores. Two other measures related to classification accuracy—false negative and false positive rates—are also reported. False positive happens when a test taker is falsely assigned to a proficiency level that is higher than their true proficiency. False negative occurs when a test taker is falsely assigned to a proficiency level that is below their true proficiency.

Table 7. Decision Accuracy and Consistency: CELPIP Listening Scores

		ACCURACY		
CELPIP LEVEL	CLASSIFICATION CONSISTENCY	Classification Accuracy	False Negative	False Positive
4	0.96	0.97	0.02	0.01
5	0.94	0.96	0.02	0.02
6	0.92	0.95	0.03	0.02
7	0.90	0.93	0.04	0.03
8	0.87	0.91	0.05	0.04
9	0.85	0.89	0.06	0.05
10	0.85	0.89	0.06	0.05

Table 8. Decision Accuracy and Consistency: CELPIP Reading Scores

		ACCURACY		
CELPIP LEVEL	**CLASSIFICATION CONSISTENCY**	Classification Accuracy	False Negative	False Positive
4	0.96	0.97	0.02	0.01
5	0.94	0.96	0.02	0.02
6	0.92	0.94	0.03	0.02
7	0.90	0.93	0.04	0.03
8	0.88	0.91	0.05	0.04
9	0.87	0.91	0.05	0.04
10	0.87	0.90	0.05	0.05

Table 9. Decision Accuracy and Consistency: CELPIP Writing Scores

		ACCURACY		
CELPIP LEVEL	**CLASSIFICATION CONSISTENCY**	Classification Accuracy	False Negative	False Positive
4	0.99	0.99	0.00	0.00
5	0.97	0.98	0.01	0.01
6	0.93	0.95	0.03	0.02
7	0.88	0.92	0.05	0.03
8	0.87	0.91	0.06	0.03
9	0.94	0.96	0.03	0.01
10	0.97	0.98	0.02	0.00

Table 10. Decision Accuracy and Consistency: CELPIP Speaking Scores

		ACCURACY		
CELPIP LEVEL	**CLASSIFICATION CONSISTENCY**	Classification Accuracy	False Negative	False Positive
4	0.98	0.98	0.00	0.02
5	0.90	0.92	0.06	0.02
6	0.93	0.95	0.04	0.01
7	0.94	0.96	0.03	0.02
8	0.95	0.97	0.02	0.01
9	0.97	0.98	0.01	0.01
10	0.97	0.98	0.01	0.01

The results above show that the classification accuracy and consistency for the four components is good, with high values for accuracy and consistency through CELPIP 4 to 10 (correspond to CLB 4–10). The false positive and false negative rates at each level are low (under 0.06).

4.3. The Conditional Standard Error of Measurement by Component

Another approach to examining scoring consistency is to estimate the conditional standard error of measurement (CSEM) for each of the CELPIP levels 4 through 10. For listening and reading components, CSEM for each level was the inverse of test information function (TIF) at the cut score. Every test form has a slightly different TIF. To be consistent with the reporting practice in Section 3, for IRT-based form statistics, the results are reported based on the analysis of the top 20 common forms.

Table 11. Conditional Standard Error of Measurement by Components

CELPIP LEVEL	LISTENING*	READING*	WRITING	SPEAKING
4	0.31	0.30	0.22	0.24
5	0.29	0.29	0.24	0.25
6	0.30	0.29	0.25	0.26
7	0.31	0.30	0.25	0.26
8	0.33	0.32	0.24	0.25
9	0.38	0.34	0.22	0.24
10	0.41	0.37	0.20	0.21

Note: *The conditional standard errors of measurement for listening and reading scores are the average of the top 20 most commonly used forms.

As shown in the table above, the CSEM is fairly low across all four components and CELPIP levels.

5. Correlations among the Four Components

Correlations across the four component scores (CELPIP levels) for the CELPIP Test are presented below in Table 12 and Table 13. Those in Table 12 are conventional Pearson correlation coefficients calculated based on continuous scores (i.e., equated true scores for reading and listening and final ratings for speaking and writing). As recommended in the psychometric research literature (Drasgow, 1986), recognizing the nature of the band scores is ordinal, the polychoric correlations to quantify the relationships between ordinal variables (see Table 13) are also reported. As illustrated in Table 12, all the polychoric correlations based on CELPIP band levels are in the range of 0.75 to 0.80, with the highest correlation observed between listening and reading (r = 0.85). Table 13 presents the results of Pearson correlations, ranging from 0.66 to 0.85. Overall, this correlation pattern suggests that the four component scores of CELPIP are positively correlated but not highly overlapped.

Table 12. Pearson Correlations among the Four Components of CELPIP (Continuous Scores)

	LISTENING	READING	WRITING
READING	0.85		
WRITING	0.74	0.76	
SPEAKING	0.66	0.66	0.74

Table 13. Polychoric Correlations among the Four Components of CELPIP (Band Levels)

	LISTENING	READING	WRITING
READING	0.85		
WRITING	0.77	0.80	
SPEAKING	0.76	0.75	0.79

6. Differential Item Functioning

Differential item functioning (DIF) is one technique used to help ensure the fairness of tests. DIF occurs when test takers of equal ability, while belonging to distinct subpopulations (e.g., male or female) perform in detectably different ways on a test item (Holland & Wainer, 1993). Paragon has used the generalized linear regression model methods for DIF detection because, as will be demonstrated below, these methods allow a common statistical framework for the varied DIF questions (e.g., Gadermann et al., 2018; Zumbo, 2007a, 2008).

Regression-based DIF methods are chosen because of their flexibility. They could detect both uniform and non-uniform DIF (French & Miller, 1996; Swaminathan & Rogers, 1990; Zumbo, 1999, 2008), allow either observed or latent variable to serve as the matching variable, and could model binary, ordinal, or continuous scores. In addition, under this framework, researchers could obtain both hypothesis significance testing results and effect size estimates. The effect size measures the magnitude of a DIF effect and aid in the interpretability of the results.

Effect size measures are particularly useful when analyses are conducted with large sample sizes, such as those in operational testing programs such as those found at Paragon, wherein the statistical power is so great as to flag even very small statistical effects. Paragon has adopted the widely used A, B, or C classification of DIF outcomes (Zieky, 2003; Zwick, 2012) adapted for the regression-based DIF methods (Jodoin & Gierl, 2001). As described below, the category into which a question will be placed depends on two factors: both statistical significance and the magnitude of the effect size. Based on over three decades of use in operational testing (Zieky, 2003; Zumbo, 2008), items are designated as category A (negligible or nonsignificant DIF), B (slight to moderate DIF), or C (moderate to large DIF). Category B items tend to have minimal impact on test scores with tests the length of those used at Paragon (Zieky, 2003; Zumbo, 2008). As Zieky notes, in typical operational testing contexts category C items are referred to item writers for a close inspection and review and would be used in operational form assembly only when the items are essential to meet important test specifications and no alternative item is available in the item bank. To date, Paragon tends to take category C items out of operational testing.

6.1. Gender DIF

For listening and reading test items, binary logistic regression models were used to detect DIF between gender groups (males vs. females). Items were classified into three categories by Jodoin and Gierl's (2001) effect size criteria: category A, *negligible or nonsignificant* (change in Nagelkerke $R^2 < 0.035$); category B, *moderate* (change in Nagelkerke R^2 between 0.035 and 0.070); or category C, *large* (change in Nagelkerke $R^2 > 0.070$).

A total number of 2,921 listening items and 2,459 reading items met the minimal sample size requirement. More specifically, to be included in gender DIF analysis, an item needs to have responses from at least 200 test takers from each group. Table 14 summarizes the gender DIF results.

Table 14. Gender DIF Results for Listening and Reading

	GENDER DIF RESULTS			ITEMS INVESTIGATED
COMPONENT	No DIF (A)	Slight to moderate DIF (B)	Moderate to large DIF (C)	
LISTENING	2918	2	1	2921
READING	2457	2	0	2459

As shown by the DIF results, nearly all (2918 out of 2921 = 99.9%) of the listening and reading items did not display DIF, that is, they function consistently between gender groups. Among the three listening items that were flagged showing gender DIF, one (category B, uniform DIF) slightly favors male test takers, one (category B, uniform DIF) slightly favors female test takers, and one (category C, non-uniform DIF) favors males at mid-to-high proficiency and it favors females at low proficiency. No reading items were flagged as category C gender DIF; two out of 2,459 items were classified as category B uniform DIF with one favoring male test takers and the other favoring females. To summarize, it is unlikely that the test scores obtained on the CELPIP listening and reading components are dependent on test-takers' gender.

A similar approach was employed to identify CELPIP writing and speaking prompts exhibiting DIF between sub-populations. Prompts to which test takers from different genders (males vs. females) with the same functional language ability may have achieved different scores were sought. Writing and speaking prompts which function differently across groups may negatively affect the comparability of test score across sub-populations and thus, raise fairness concerns. Paragon regularly monitors item performance to ensure any items with security or quality concerns are not administered to test takers.

For this report, all the writing and speaking prompts that were analyzed have been used for operational purposes between March 1, 2017 and March 1, 2020. To be included in the DIF analysis, the writing and speaking prompts have meet the minimum sample size requirement of 200 test takers within each group. DIF analysis was conducted for a total of 288 writing prompts and 1041 speaking prompts.

The rule adopted to flag prompts showing DIF was a combined rule, where (1) The F test between model 3 and model 1 is significant, suggesting a significant improvement of the overall model by including grouping variable and the possible interaction terms; and (2) the change of R^2 between model 3 and model 1 falls into 0.035 to 0.070 (moderate DIF) or it is larger than 0.070 (large DIF) (adapted from Jodoin & Gierl, 2001).

Among the 288 writing and 1041 speaking prompts investigated, none of them was flagged as showing gender DIF. That is, all of the writing and speaking prompts investigated fall into category A (negligible or nonsignificant DIF).

6.2. Language Group DIF

DIF due to test-takers' self-reported first language (L1) groups is also investigated and reported below. The analytical strategy is consistent with the methods used for examining gender DIF. Different from the situation when investigating gender DIF where the number of groups is small and fixed, i.e., the comparison is always between two groups (males and females), a large number of possible first language groups exist (more than 100 first language groups are reported by CELPIP test takers). To analyze multiple language groups simultaneously, Paragon uses mixed effects models. More specifically, listening and reading items were investigated using a generalized linear mixed model (GLMM) and writing and speaking prompts were examined using linear mixed model. This approach is in line with the view of exchangeability described in Zumbo (2007b).

To interpret the L1 DIF effect, an intraclass correlation coefficient (ICC) was calculated for each investigated item. The ICC is used to quantify the proportion of variance that could be attributed to the variation across L1 groups:

$$\text{ICC} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_f^2 + \sigma_\varepsilon^2}$$

where σ_r^2 is the variance component associated with the random effect L1 groups, $\sigma_f^2$is the variance component attributes to the fixed effect of matching variable, and σ_ε^2 is the residual variance.

For the ease of interpretation, the Zumbo and Thomas scale (Zumbo & Thomas, 1997; Zumbo, 2008) was adopted to classify each L1 DIF effect into one of the three categories based on their ICC value. Specifically, "A" means negligible DIF effect (ICC < 0.13), "B" means moderate DIF effect (ICC between 0.13 and 0.26), and "C" means large DIF effect (ICC > 0.26).

For an L1 group to be included in the DIF analysis, its sample size needs to be at least 50. In other words, for each item under investigation, at least 50 test takers from the same L1 background need to have responded to it to allow them to be included in the DIF analysis. In total, 2552 listening items, 2072 reading items, 229 writing prompts, and 911 speaking prompts were analyzed.

An example of the statistics that are collected to evaluate each item for language DIF is presented in Table 15. As described above, the main interest is the value of ICC for each item and the Zumbo and Thomas scale (denoted as ZT in the table) was applied to interpret the magnitude of the DIF effect. Results for L1 DIF are summarized in Table 16 below.

Table 15. Language Background DIF Example

ITEM CODE	FIXED INTERCEPT EFFECT	RANDOM INTERCEPT VARIANCE COMPONENT	ICC	ZT	L1 GROUPS INVESTIGATED FOR THIS ITEM	TOTAL SAMPLE SIZE
R1_0072_05	-2.36	0.78	0.15	B	17	3241

Note: ICC denotes the intra-class correlation, ZT is the Zumbo-Thomas scale, and L1 standards for self-report first language.

Table 16. Summary of the L1 Background DIF Results

	L1 BACKGROUND DIF RESULTS			**ITEMS INVESTIGATED**
COMPONENT	No DIF (A)	Slight to moderate DIF (B)	Moderate to large DIF (C)	
LISTENING	2536	15	1	2552
READING	2058	13	1	2072
WRITING	229	0	0	229
SPEAKING	911	0	0	911

As shown in the summary table above, none of the writing or speaking prompts was flagged as showing DIF due to L1 backgrounds. Overall, 2 out of 4624 (about 0.04%) listening and reading items were flagged as category C, only 28 (about 0.61%) were flagged as category B, and all the remaining 4594 (99.35%) fall into category A (no DIF). Thus, it is highly *unlikely* that the test scores be unduly affected by test takers' L1 backgrounds. As a reminder, Paragon takes category C items out of operational testing.

6.3. Drift

In this section, an analysis of item parameter drift and rating drift over the previous three years of CELPIP administrations are presented.

6.3.1. Item Parameter Drift over Time

Drift is likely to occur when maintaining an item pool over time even though good quality items are selected and test content is secured carefully. Item drift (or item parameter drift) may be expected because of frequent item exposure or test-takers' pre-knowledge of the test content. Items may also perform differently across years due to changes in the test structure and content or test-taker population.

Item drift is monitored by examining the changes in item difficulty over continuing use. To investigate item drift, Paragon uses the statistical method of differential item functioning (DIF). DIF methods investigate whether different groups of test takers (e.g., males and females), who have the same measured ability, have different probabilities of achieving the same score on one item.

When the DIF technique is applied to examine item drift over time, test takers at different time points are regarded as different groups. Specifically, for this report, test takers were divided into three groups based on the time they took the test, i.e., March 1, 2017 - February 28, 2018; March 1, 2018 - February 28, 2019; and March 1, 2019 - March 1, 2020. Scoring drift is then investigated by examining the time-group differences in the likelihood of achieving the same item-level score after controlling for test candidates' proficiency levels.

Consistent with the DIF methodology adopted by Paragon, item drift was investigated using the regression models.

Drift items are flagged by evaluating the change of R^2 from baseline model to uniform DIF model. To mimic the common procedures used for DIF investigation, the items were classified into three categories of drift, A, B, and C using Jodoin and Gierl's (2001) effect size criteria. According to their criteria, category A shows negligible effect (change in Nagelkerke $R^2 < 0.035$); category B has moderate effect (change in Nagelkerke R^2 between 0.035 and 0.070); and category C exhibits large effect (change in Nagelkerke $R^2 > 0.070$). For writing and speaking prompts, where the raw rating scores are on a continuous scale of 0 to 12, linear regression models were employed and thus, R^2 for multiple regressions rather than Nagelkerke R^2 were used for logistic regression models as effect size measures. Similar to common practices in dealing with DIF items, items categorized as category A or B are viewed as showing negligible to small item difficulty drift and items categorized as category C is marked for further review.

In total, 1267 listening items, 1115 reading items, 59 writing prompts, and 304 speaking prompts met Paragon's inclusion criteria and were investigated for item drift. To be included in the analysis, these items and prompts (1) have been used at the three defined time periods and (2) within each time period, they were administered to a minimum of 200 test takers. None of the items or prompts have been flagged as showing drift (category B or C), suggesting these operational items have functioned in similar ways across the investigated time periods. That is, all items fall into category A (nonsignificant or negligible) drift.

6.3.2. Rating Drift

In addition to item drift, speaking and writing scores that are evaluated by raters are also subject to changes in raters' behavior. Changes in raters' behaviour is called rater drift, and occurs when raters unintentionally redefine their scoring criteria or standards over time (Wheeler, Haertel, & Scriven, 1992, p. 12). Despite attempts to maintain constant standards, rating drift may still happen due to increasing practice effect, inconsistent ratings within a rater (e.g., non-adherence to the scoring rubric or rating procedures), and changes to the rater training, the prompt/test, and test-taker population (e.g., Congdon & McQueen, 2000; McKinley & Boulet, 2004; Myford & Wolfe, 2003).

One of the most prevalent effects of rater drift is the rater-severity effect. This effect occurs when raters provide ratings that are consistently too harsh or too lenient, as compared to other raters. Severity effects can be explicitly modeled in a multifaceted Rasch model (MFRM) framework (Linacre, 1989), and thus, to evaluate rater drift (i.e., change in severity), for speaking and writing raters separately, Paragon conducted three MFRMs for the following three time periods, respectively: March 1, 2017 - February 28, 2018; March 1, 2018 - February 27, 2019; and February 28, 2019 - March 1, 2020.

The MFRM analyses were performed using Facet (Linacre, 2011). For each MFRM, all the active raters during the time period were included in the modeling process. The rater-severity measures were retrieved from three independent MFRMs and then equated to the scale of the first model (i.e., the model using data from March 1, 2017 to February 28, 2018) using the mean-sigma method (Marco, 1977). A total of 57 writing raters and 80 speaking raters received their severity measures for each of the three periods of time, and thus, were included in the subsequent comparisons to examine the rater drift effect. Each of these writing raters evaluated more than 400 assignments during each time period; each of these speaking raters completed more than 250 assignments during each time period.

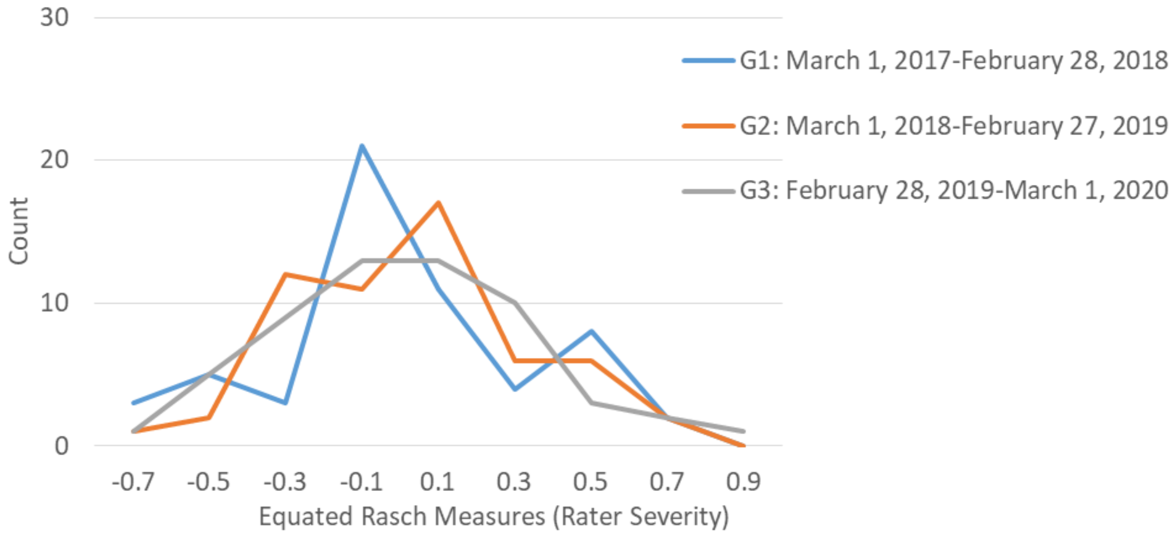

Figure 3. The Distributions of Rater Severity (Writing Raters)

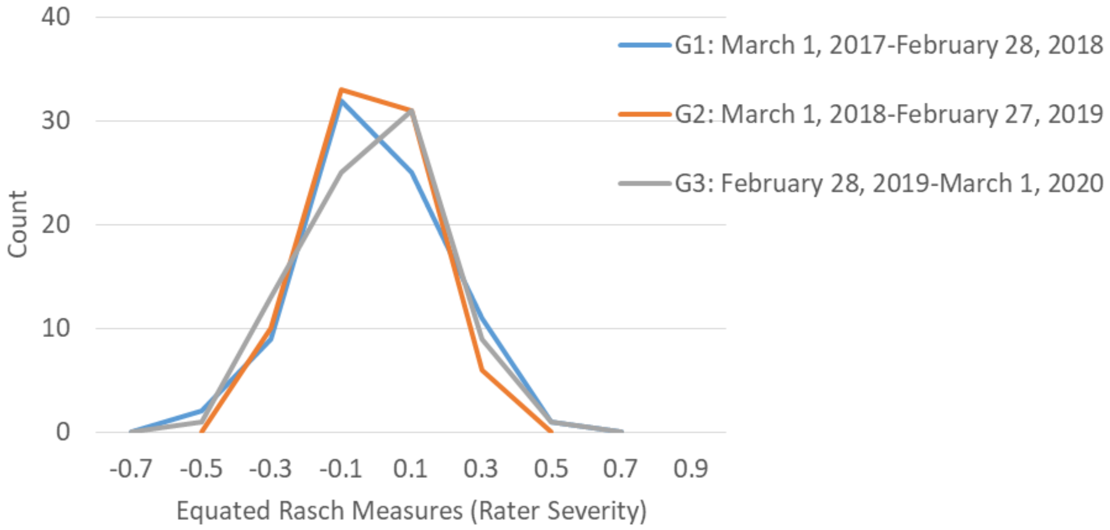

Figure 4. The Distributions of Rater Severity (Speaking Raters)

As shown by the figures above, the distributions of the rater-severity measures largely overlap with each other. They suggest that over time, the overall severity of all the raters included in the analysis was consistent across the three periods of time from March 1, 2017 to March 1, 2020. The results presented in Table 17 showing the summary statistics of raters' changes in their severity, further confirm this observation.

Table 17. Changes of Rater Measures over Time

	G1 VS. G2		**G1 VS. G3**	
COMPONENT	mean	Standard deviation	mean	Standard deviation
WRITING RATERS	0.033	0.246	-0.003	0.329
SPEAKING RATERS	0.002	0.130	0.002	0.192

Note: G1: March 1, 2017-February 28, 2018; G2: March 1, 2018-February 27, 2019; and G3: February 28, 2019-March 1, 2020

7. Concluding Remarks

This report has presented psychometric analyses of the CELPIP Test to demonstrate the consistency and accuracy of the test in classifying test takers across proficiency levels and administrations. Paragon did not identify any substantive concerns in any set of analysis. Specifically, the results suggest the following:

- The Classical Test Theory (CTT) statistics were summarized for forms with a minimum of 200 test takers, and the IRT statistics for listening and reading components were presented for the most common 20 forms. Reliability for each component by sub-forms is high. Distributions of test information functions (TIFs) are similar across different sub-forms for both listening and reading components.
- Decision consistency and decision accuracy for CELPIP Levels 4 to 10 (which correspond to CLB levels 4 to 10) are high for all components.
- Differential item functioning analyses have identified no concern in terms of items or writing or speaking prompts that could be functioning differently between self-reported gender or first language groups for all four components.
- Item parameter drift analyses have identified no items or writing or speaking prompts as showing significant drift, suggesting these operational items have functioned in similar ways across the investigated time periods.
- Rater drift analysis shows the overall severity of the raters included in the analysis was consistent across the three periods of time from March 1, 2017 to March 1, 2020.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Brennan, R. L. (2004). *BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy* (Version 1.0) (CASMA Research Report No. 9). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment at the University of Iowa.

Cai, L. (2017). *flexMIRT®: Flexible multilevel multidimensional item analysis and test scoring*. [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163-178.

Drasgow, F. (1986). Polychoric and polyserial correlations. In Kotz, Samuel, Narayanaswamy Balakrishnan, Campbell B. Read, Brani Vidakovic & Norman L. Johnson (Eds), *Encyclopedia of Statistical Sciences, Vol. 7,* (pp. 68-74). New York, NY: John Wiley.

French, A.W., & Miller, T. R., (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*, 315–332.

Gadermann, A. M., Chen, M. Y., Emerson, S. D., & Zumbo, B. D. (2018). Examining validity evidence of self-report measures using differential item functioning: An illustration of three methods. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 14*(4), 164–175. https://doi.org/10.1027/1614-2241/a000156

Holland, P. W., & Wainer, H. Eds, (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression for DIF detection. *Applied Measurement in Education, 14*(4), 329–49.

Lathrop, Q. N. (2014). R Package cacIRT: Estimation of classification accuracy and consistency under item response theory. *Applied Psychological Measurement, 38*(7), 581-582.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (2011). *A user's guide to FACETS: Rasch-model computer programs*. Chicago, IL: Winsteps.com

Livingston, S. A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179-197.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*(2), 139-160.

McKinley, D. W., & Boulet, J. R. (2004). Detecting score drift in a high-states performance-based assessment. *Advances in Health Sciences Education, 9*, 29-38.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386-422.

Rudner, L. M. (2001) Computing the expected proportions of misclassified examinees. *Practical Assessment, Research, & Evaluation, 7*(14), 1-5.

Rudner, L. M. (2005) Expected classification accuracy. *Practical Assessment Research & Evaluation, 10*(13), 1-4.

Swaminathan, H., & Rogers, H. (1990). Detecting differential item functioning using logistic regression procedures, *Journal of Educational Measurement, 27*, 361–370.

Wheeler, P., Haertel, G., & Scriven, M. (1992). *Teacher evaluation glossary*, Kalamazoo, MI: CREATE Project, The Evaluation Center, Western Michigan University.

Zieky, M. (2003). *A DIF Primer*. Princeton, NJ: ETS.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2007a). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233.

Zumbo, B. D. (2007b). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 45–79). Amsterdam, The Netherlands: Elsevier Science.

Zumbo, B. D. (2008). *Statistical methods for investigating item bias in self-report measures [The University of Florence Lectures on Differential Item Functioning]*. Universita degli Studi di Firenze, Florence, Italy.

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF*. Prince George, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioural Science.

Zwick, R. (2012). *A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement*. Princeton, NJ: ETS.